Guide de Prompting GPT-4.1

Les Points Clés du Guide Officiel d'OpenAI

Document de formation préparé par VISION IA

Introduction à GPT-4.1

GPT-4.1 représente la dernière évolution des modèles d'intelligence artificielle d'OpenAI, avec des améliorations significatives par rapport à ses prédécesseurs :

Caractéristiques distinctives :

- Plus rigoureux dans ses réponses
- Plus obéissant aux instructions spécifiques
- Plus précis dans ses analyses
- Moins permissif dans ses interprétations

La nouvelle version suit les instructions au mot près et n'infère plus forcément vos intentions comme le faisaient les versions précédentes. Sans adapter vos prompts, vous n'exploiterez qu'une fraction de son potentiel réel.

💡 À retenir : GPT-4.1 est une machine de guerre, mais si vous continuez à formuler vos prompts comme en 2023, vous n'exploitez pas même 30% de sa puissance.

Mode Agent (Prompting Agentique)

Un concept-clé partagé par OpenAI est le "mode agent" - une façon d'écrire vos prompts pour que GPT ne se contente pas de répondre partiellement, mais qu'il aille au bout du problème, comme un véritable partenaire de travail.

Selon le guide officiel, trois éléments sont essentiels pour transformer le comportement de GPT-4.1 :

1. Persistance

Indiquez clairement que l'IA doit persévérer jusqu'à la résolution complète du problème.

```
"Continue jusqu'à ce que ce soit entièrement terminé. Ne t'arrête pas tant que ce n'est pas 100% résolu."
```

Dans leur exemple, OpenAI recommande :

```
"Ne termine ton tour que lorsque tu es certain que le problème est résolu."
```

2. Utilisation des Outils Mentaux

Encouragez l'IA à réfléchir plutôt qu'à deviner ou approximer :

```
"Si tu n'es pas sûr, ne devine pas. Pose-toi des questions, explique ton raisonnement étape par étape."
```

OpenAI insiste :

```
"Ne devine PAS et n'invente PAS de réponse."
```

3. Planification

Demandez à l'IA d'élaborer un plan avant de répondre :

```
"Avant de répondre, fais un plan. Réfléchis à haute voix. Dis-moi comment tu comptes t'y prendre."
```

OpenAI confirme que cette approche améliore significativement la qualité des réponses :

```
"Tu DOIS planifier minutieusement chaque action, puis réfléchir aux résultats obtenus."
```

📈 Résultat : OpenAI note que ces trois instructions ont amélioré les performances d'environ 20% lors de leurs tests.

Gestion du Contexte Long (1 million de tokens)

Une des avancées majeures de GPT-4.1 est sa capacité à gérer des contextes immenses - jusqu'à 1 million de tokens, soit l'équivalent de centaines de pages de contenu dans une seule conversation.

Capacités avec le contexte long :

- Analyse structurée de documents entiers
- Re-classement des informations pertinentes
- Sélection intelligente des éléments utiles
- Filtrage du contenu non pertinent

Techniques pour l'organisation du contexte

⚠ Astuce cruciale : "Placez vos instructions à la fois au début ET à la fin du contexte fourni."

Pourquoi ? Dans un contexte très long, GPT peut perdre de vue ce qu'on lui a demandé au début. La solution est d'encadrer votre prompt comme un sandwich :

Structure recommandée :
INSTRUCTION → CONTEXTE → INSTRUCTION

Structuration du contenu

Pour les blocs de texte très longs, utilisez des balises claires :

Type de balise	Exemple
Titres Markdown	### Partie 1 – Résumé ### Partie 2 – Analyse
Balises XML	<section>...</section>
Séparateurs visuels	--- ou >>>

💡 À retenir : La façon dont vous structurez le prompt impacte directement la qualité des réponses, surtout avec un contexte long.

Suivi Littéral des Instructions

Une différence notable avec GPT-4.1 : il ne lit plus entre les lignes. Il suit les consignes à la lettre, parfois au détriment de votre intention initiale.

⚠ Avertissement d'OpenAI : "GPT-4.1 suit les instructions de manière plus littérale. Les prompts existants peuvent ne pas fonctionner comme prévu."

Avant (GPT-4)

Pouvait être flou, suggérer, insinuer...
Le modèle devinait souvent ce que vous vouliez dire.

Maintenant (GPT-4.1)

Besoin de consignes explicites, précises, détaillées.
Le modèle fait exactement ce que vous écrivez.

Structure de Prompt Recommandée

OpenAI propose une méthode simple mais puissante pour structurer vos prompts en 4 blocs :

1. # RÈGLES
Principes généraux : quel ton ? quel rôle ? quoi éviter ?

```
"Tu es un expert en psychologie. Sois clair, bienveillant, et évite les termes trop techniques."
```

2. # ÉTAPES
Méthode à suivre : les étapes, leur ordre, les vérifications

```
"1. Résume le texte.
2. Identifie les émotions principales.
3. Donne des conseils pratiques adaptés à chaque émotion."
```

3. # EXEMPLES
Un ou deux exemples concrets de ce que vous attendez
Attention : trop d'exemples tue la créativité. Un ou deux bien choisis suffisent.

4. # FORMAT
Présentation souhaitée du résultat final

Erreurs à Éviter

Erreur	Pourquoi éviter
Majuscules à outrance	"TU DOIS TOUJOURS FAIRE ÇA !!!" n'est pas plus efficace et peut rendre GPT-4.1 trop rigide.
Règles contradictoires	Si vous dites "fais court" puis "explique en détail", GPT suivra la dernière consigne.
Consignes floues	'Fais un truc stylé' ou 'sois plus humain' ne sont pas assez précis pour un modèle.

⚠ Rappel important : GPT-4.1 n'est pas un lecteur de pensées. Il exécute vos instructions au pied de la lettre.

Exemple concret

Prompt flou :

"Écris un message pour un client qui s'est plaint."

→ Résultat : réponse générique, peu personnalisée

Prompt structuré :

RÈGLES
Tu es un assistant commercial dans une entreprise de logiciels. Tu réponds avec empathie, sans accuser le client.
ÉTAPES
1. Remercier pour le retour.
2. Reconnaître la gêne.
3. Proposer une solution.
4. Terminer sur une note positive.
EXEMPLE
"Bonjour, merci pour votre retour. Nous comprenons votre frustration..."
FORMAT
Texte court, 4 à 5 lignes max.

→ Résultat : message pro, humain, structuré

Chain of Thought (Chaîne de Pensée)

GPT-4.1 n'est pas intrinsèquement un modèle de raisonnement, mais on peut le forcer à raisonner avec la technique de la "Chaîne de Pensée" (Chain of Thought).

Cette méthode consiste à obliger le modèle à "penser à voix haute" plutôt que de simplement donner une réponse directe.

Dans leurs tests sur le benchmark SWE-bench Verified, OpenAI a constaté que l'ajout d'un raisonnement structuré a augmenté les performances de 4% – un gain significatif pour un benchmark technique. Pour les cas d'usage quotidiens, le gain peut être de 10 à 20%.

Structure de Raisonnement Avancée

OpenAI propose une "Stratégie de Raisonnement" complète :

```
# Stratégie
   Analyse de la question : que veut réellement l'utilisateur ?
   Analyse du contexte : quels documents ou infos sont pertinents ?
      Pour chaque élément : est-ce utile ? Pas utile ?
   Synthèse : résumer les documents les plus pertinents, puis répondre.
```

Et le prompt se termine généralement par :

```
"Réfléchis d'abord étape par étape à ce qu'il faut faire, puis donne ta réponse finale."
```

Synthèse des 5 Leviers Essentiels

Levier	Description
Prompting Agentique	Transformer GPT en assistant autonome via persistance, outils mentaux et planification
Utilisation des Outils	Comprendre la logique d'utilisation des outils pour inspirer vos propres cas d'usage
Contexte Long	Exploiter jusqu'à 1 million de tokens avec une structure de prompt adaptée
Suivi Littéral	Donner des instructions claires, précises et structurées que GPT-4.1 suivra à la lettre
Chaîne de Pensée	Pousser l'IA à penser à voix haute pour améliorer significativement ses réponses

Créé avec Genspark

💡 **Conseil final :** Testez ces techniques une par une pour comprendre leur impact individuel, puis combinez-les pour des résultats optimaux avec GPT-4.1.